Attorney Docket No.: 0160116

UNITED STATES PATENT APPLICATION

FOR

RECOVERING AN ERASED VOICE FRAME WITH TIME WARPING

INVENTORS:

EYAL SHLOMOT
YANG GAO

CERTIFICATE OF EXPRESS MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service "Express Mail Post Office to addressee" Service under 37 C.F.R. Sec. 1.10 addressed to: Commissioner for Patents, P. O. Box 1450, Alexandria, VA 22313-1450, on 3/11/04

Express Mailing Label No.: EV420421887US

Lori Lapidario — Lori Lapidario

Name — Signature

PREPARED BY:

FARJAMI & FARJAMI LLP
26522 La Alameda Ave., Suite 360
Mission Viejo, California 92691

(949) 282-1000
Customer No. 25700

25700
PATENT TRADEMARK OFFICE

RECOVERING AN ERASED VOICE FRAME WITH TIME WARPING

RELATED APPLICATIONS

The present application claims the benefit of United States provisional application serial number 60/455,435, filed March 15, 2003, which is hereby fully incorporated by reference in the present application.

United States Patent Application Serial Number _____, "SIGNAL DECOMPOSITION OF VOICED SPEECH FOR CELP SPEECH CODING," Attorney Docket Number: 0160112.

United States Patent Application Serial Number _____, "VOICING INDEX CONTROLS FOR CELP SPEECH CODING," Attorney Docket Number: 0160113.

United States Patent Application Serial Number _____, "SIMPLE NOISE SUPPRESSION MODEL," Attorney Docket Number: 0160114.

United States Patent Application Serial Number _____, "ADAPTIVE CORRELATION WINDOW FOR OPEN-LOOP PITCH," Attorney Docket Number: 0160115.

BACKGROUND OF THE INVENTION

1. FIELD OF THE INVENTION

The present invention relates generally to speech coding and, more particularly, to recovery of erased voice frames during speech decoding.

2. RELATED ART

From time immemorial, it has been desirable to communicate between a speaker at one point and a listener at another point. Hence, the invention of various

telecommunication systems. The audible range (i.e. frequency) that can be transmitted and faithfully reproduced depends on the medium of transmission and other factors. Generally, a speech signal can be band-limited to about 10 kHz without affecting its perception. However, in telecommunications, the speech signal bandwidth is usually limited much more severely. For instance, the telephone network limits the bandwidth of the speech signal to between 300 Hz to 3400 Hz, which is known in the art as the "narrowband". Such band-limitation results in the characteristic sound of telephone speech. Both the lower limit at 300Hz and the upper limit at 3400 Hz affect the speech quality.

In most digital speech coders, the speech signal is sampled at 8 kHz, resulting in a maximum signal bandwidth of 4 kHz. In practice, however, the signal is usually band-limited to about 3600 Hz at the high-end. At the low-end, the cut-off frequency is usually between 50 Hz and 200 Hz. The narrowband speech signal, which requires a sampling frequency of 8 kb/s, provides a speech quality referred to as toll quality. Although this toll quality is sufficient for telephone communications, for emerging applications such as teleconferencing, multimedia services and high-definition television, an improved quality is necessary.

The communications quality can be improved for such applications by increasing the bandwidth. For example, by increasing the sampling frequency to 16 kHz, a wider bandwidth, ranging from 50 Hz to about 7000 Hz can be accommodated. This bandwidth range is referred to as the "wideband". Extending the lower frequency range to 50 Hz increases naturalness, presence and comfort. At the other end of the spectrum, extending the higher frequency range to 7000 Hz increases intelligibility and makes it easier to differentiate between fricative sounds.

The frame may be lost because of communication channel problems that results

in a bitstream or a bit package of the coded speech being lost or destroyed. When this happens, the decoder must try to recover the speech from available information in order to minimize the impact on the perceptual quality of speech being reproduced.

Pitch lag is one of the most important parameters for voiced speech, because the perceptual quality is very sensitive to pitch lag. To maintain good perceptual quality, it is important to properly recover the pitch track at the decoder. Thus, a traditional practice is that if the current voiced frame bitstream is lost, pitch lag is copied from the previous frame and the periodic signal is constructed in terms of the estimated pitch track. However, if the next frame is properly received, there is a potential for quality impact because of discontinuity introduced by the previously lost frame.

The present invention addresses the impact in perceptual quality due to discontinuities produced by lost frames.

SUMMARY OF THE INVENTION

In accordance with the purpose of the present invention as broadly described herein, there is provided systems and methods for recovering an erased voice frame to minimize degradation in perceptual quality of synthesized speech.

In one embodiment, the decoder reconstructs the lost frame using the pitch track from the directly prior frame. When the decoder receives the next frame data, it makes a copy of the reconstructed frame data and continuously time warping it and the next frame data so that the peaks of their pitch cycles coincide. Subsequently, the decoder fades out the time-warped reconstructed frame data while fading in the time-warped next frame data. Meanwhile, the endpoint of the next frame data remains fixed to preclude discontinuity with the subsequent frame.

These and other aspects of the present invention will become apparent with further reference to the drawings and specification, which follow. It is intended that all such additional systems, methods, features and advantages be included within this description, be within the scope of the present invention, and be protected by the accompanying claims.

BRIEF DESCRIPTION OF DRAWINGS

Figure 1 is an illustration of the time domain representation of a coded voiced speech signal at the encoder.

Figure 2 is an illustration of the time domain representation of the coded voiced speech signal of Figure 1, as received at the decoder.

Figure 3 is an illustration of the discontinuity in the time domain representation of the coded voiced speech signal after recovery of a lost frame.

Figure 4 is an illustration of the time warping process in accordance with an embodiment of the present invention.

Figure 5 illustrates real-time voiced frame recovery in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION

The present application may be described herein in terms of functional block components and various processing steps. It should be appreciated that such functional blocks may be realized by any number of hardware components and/or software components configured to perform the specified functions. For example, the present application may employ various integrated circuit components, e.g., memory elements, digital signal processing elements, transmitters, receivers, tone detectors, tone generators, logic elements, and the like, which may carry out a variety of functions under the control of one or more microprocessors or other control devices. Further, it should be noted that the present application may employ any number of conventional techniques for data transmission, signaling, signal processing and conditioning, tone generation and detection and the like. Such general techniques that may be known to those skilled in the art are not described in detail herein.

Figure 1 is an illustration of the time domain representation of a coded voiced speech signal at the encoder. As illustrated, the voiced speech signal is separated into frames (e.g. frames 101, 102, 103, 104, and 105) before coding. Each frame may contain any number of pitch cycles (i.e. illustrated as big mounds). Each frame is transmitted from the encoder to the receiver as a bitstream after coding. Thus, for example, frame 101 is transmitted to the receiver at t_{n-1}, frame 102 at t_n, frame 103 at t_{n+1}, frame 104 at t_{n+2}, frame 105 at t_{n+3}, and so on.

Figure 2 is an illustration of the time domain representation of the coded voiced speech signal of Figure 1, as received at the decoder. As illustrated, frame 101 arrives properly at the decoder as frame 201; Frame 103 arrives properly at the decoder as frame 203; Frame 104 arrives properly at the decoder as frame 204; and Frame 105 arrives properly at the decoder as frame 205. However, frame 102 does

not arrive at the decoder because it was lost in transmission. Thus, frame 202 is blank.

To maintain perceptual quality, frame 202 must be reproduced at the decoder in real-time. Thus frame 201 is copied into frame 202 slot as frame 201A. However, as shown in Figure 3, a discontinuity may exist at the intersection of frames 201A and 203 (i.e. point 301) because the previous pitch track (i.e. frame 201A) is likely not accurate . This is because frame 203 was properly received thus its pitch track is correct. But since frame 201A is a reproduced frame 201, its endpoint may not coincide with the beginning point of correct frame 203 thus creating a discontinuity that may affect perceptual quality.

Thus, although frame 201A is likely incorrect , it may no longer be modified since it has already been synthesized (i.e. it's time has passed and the frame has been sent out). The discontinuity at 301 created by the lost frame may produce an audible reproduction at the beginning of the next frame that is annoying.

Embodiments of the present invention use continuous time warping to minimize impact on perceptual quality. Time warping involves mainly modifying or shifting the signals to minimize the discontinuity at the beginning of the frame and also improve the perceptual quality of the frame. The process is illustrated using Figure 4 and Figure 5. As illustrated in Figure 4, time history 420 is the actual received data (see Figure 2) showing the lost frame 202. Time history 410 is a pseudo received data constructed from the received data. Time history 410 is constructed in real-time by placing a copy of received frame 201 into frame slot 202 as frame 201A and into frame slot 203 as frame 201B. Note that frame 203, frame 204, and frame 205 arrive properly in real-time and are correctly received in this illustration.

The process involves continuously time warping frames 201B of 410 and frame

203 of 420 so that their peaks, 411 and 421, coincide in time while maintaining the intersection point (e.g. endpoint 422) between frames 203 and 204 fixed. For instance, peak 411 may be stretched forward (as illustrated by arrow 414) in time by some delta while peak 421 is stretched backward (as illustrated by arrow 424) in time. The intersection point 422 must be maintained because the next frame (e.g. 204) may be a correct frame and it is desired to keep continuity between the current frame and the correct next frame, as in this illustration. After time-warping, an overlap-add of the two signals of the warped frames may be used to create the new frame. Line 413 fades out the reconstructed previous frame while line 423 fades in the current frame. The sum of curves 413 and 423 has a magnitude of one at all points in time. Figure 5 illustrates real-time voiced frame recovery in accordance with an embodiment of the present invention.

As illustrated in Figure 5, a current frame of voiced data is received in block 502. A determination is made in block 504 whether the frame is properly received. If not, the previous frame data is used to reconstruct the current frame data in block 506 and processing returns back to block 502 to receive the next frame data. If, on the other hand, the current frame data is properly received (as determined in block 504), further determination is made in block 508 whether the previous frame was lost, i.e., reconstructed. If the previous frame was not lost, the decoder proceeds to use the current frame data in block 510 and then returns back to block 502 to receive the next frame data.

If, on the other hand, the previous frame data was lost received (as determined in block 508) and the current frame data is properly received, then time warping is necessary. In block 512, the pitch of the current frame and that of the reconstructed frame is time-warped so that they will coincide. During time-warping, the end-point

of the current frame is maintained because the next frame may be a correct frame.

After the frames are time warped in block 512, the time-warped current frame is faded in while the time-warped reconstructed frame is faded out in block 514. The combined fade-in and fade-out process (over-lap-add process) may take on the form of the following equation:

$$NewFrame(n) = ReconstFrame(n) \cdot [1-a(n)] + CurrentFrame(n) \cdot a(n),$$

$$n=0, 1, 2, ..., L-1;$$

where $0<=a(n)<= 1$, usually $a(0)=0$ and $a(L-1)=1$.

After the fade process is completed in block 514, processing returns to block 502 where the decoder awaits receipt of the next frame data. Processing continues for each received frame and the perceptual quality is maintained.

The methods and systems presented above may reside in software, hardware, or firmware on the device, which can be implemented on a microprocessor, digital signal processor, application specific IC, or field programmable gate array ("FPGA"), or any combination thereof, without departing from the spirit of the invention. Furthermore, the present invention may be embodied in other specific forms without departing from its spirit or essential characteristics. The described embodiments are to be considered in all respects only as illustrative and not restrictive.